# User Perceptions of Associative Thesaural Relationships: A Preliminary Study

Rachel Ivy Clarke, University of Washington
Jin Ha Lee, University of Washington

**Abstract**
This study investigates whether potential library users can identify distinctions of associative relationships in Library of Congress Subject Headings (LCSH) and whether they perceive such relationships as useful. This work tests previous presumptions about user needs and capabilities regarding associative relationships. It lays a foundation for an interconnected network of subject relationships to navigate both library catalogs and the Web. Using sample related term pairs sourced, potential library users on Amazon Mechanical Turk were asked to identify narrower distinctions of established associative relationship types. Results indicate that some associative relationship types, especially *near-synonymous/frequently interchangeable* terms, are easily identifiable while others, like *position in time and space*, remain problematic. Despite an inclination against library catalog use, potential library users do perceive associative relationships as useful and suggest a variety of additional associative relationship types. For some users, the type of associative relationship may be less important than relationship closeness or proximity.
**Keywords:** associative relationships; Library of Congress Subject Headings; crowdsourcing; controlled vocabularies; thesauri
**Citation**: Clarke, R.I., Lee, J.A. (2015). User Perceptions of Associative Thesaural Relationships: A Preliminary Study. In *iConference 2015 Proceedings*.

**Acknowledgements**: The authors thank David Talley for his assistance in sourcing and parsing the LCSH RDF data.
**Contact**: raclarke@uw.edu, jinhalee@uw.edu

## 1 Introduction

In a progressively networked environment, relationships between entities, properties, and concepts are increasingly important. Relationship connections can help users navigate, browse, and explore a domain. Controlled vocabularies are one established way of documenting relationships. Information science scholars identified up to 120 unique relationship types in established thesauri (Association for Library Collections & Technical Services, 1999). Yet international standards stipulate the use of only three: equivalence relationships, hierarchical relationships, and associative relationships. This supposedly reduces overloading thesauri with "valueless" relationships (Aitchison, Gilchrist & Bawden, 2004) that overwhelm indexers and searchers and offer little return on intellectual, time, or labor investment (Soergel, 1974). Such assumptions have guided vocabulary construction, indexing, and thesaural retrieval research for over three decades. However, few studies have tested these assumptions, which may be increasingly relevant in a world of relational databases, semantic web technologies, and "big data" analysis. What, if any, associative thesaural relationships are a) distinguishable to users and b) potentially useful to information seekers? Improving understanding in this area can help evaluate controlled vocabularies used in the context of libraries, while also offering possible implications for the semantic web and ontologies.

This preliminary study uses associative relationships in Library of Congress Subject Headings (LCSH) to investigate if potential library users can identify narrower distinctions of established associative relationship types as well as whether or not they perceive such relationships to be useful. In addition to confirming or refuting established assumptions about associative relationship types, this study may also help provide insight to vocabulary designers, other practitioners, and scholars. Developing tools in a world increasingly driven by information links requires a full understanding of the variety of possible relationships between entities like subjects in order to assist users searching and browsing for materials in libraries or on the web at large.

## 2 Relevant Work

In information science, the definition of associative relationships can be traced back to Soergel (1974):

> Concept A is related to concept B (has an associative relationship to concept B) if the following holds: an indexer or searcher weighing the use of A should be reminded of the existence of B (and there is no hierarchical relationship between A and B).

While such an open-ended definition of association allows for contextual applications and individual indexer judgment, it is also potentially too broad to be helpful. Soergel himself describes two kinds of associative relationships: "concepts similar in meaning" and "concepts connected empirically"—the latter including terms that frequently co-occur and/or terms connected based on empirical knowledge (p. 107-109). Such descriptions are offered as guidance for vocabulary authors as a means of determining, creating, and labeling related terms (RTs) during thesaurus construction, rather than tools for end users of the thesauri—neither indexers nor searchers. However, the presence of these descriptions does indicate varied types of associative relationships.

In addition to Soregel, many scholars attempted to identify, label, and create taxonomies of relationship types. Willetts (1975) analyzed 10 contemporary thesauri, specifically examining associative relationships and rules governing their use. Despite providing some benefits, she found associative relationships were poorly defined and understood in the sample thesauri. A matrix based on conceptual categories combined with explicit relations offered 45 possible associative relationship types. Analysis of extant term pairs in the sample thesauri indicated that the type of relation is more important than the conceptual category, and that the most commonly occurring relations were appurtenance ("parts of" relationships, which may be considered hierarchical); concurrence ("mere mental juxtaposition of two concepts"); dimensional (such as spatial and temporal relations) and a general association. This general "catch-all" category was one of the most used relationships, reflecting a lack of definitions, guidelines, and consistency in the creation and application of thesaural relationships.

Neelameghan et al. (1978) described a typology of 39 non-hierarchical relationship types, using facet analysis to describe relationships based on frequently occurring facet juxtaposition. Nutter (1989) identified over 100 types of lexical relationships; however, the 15 consulted sources consisted of dictionaries and other seminal works on semantic relationships, not actual thesauri. The identified relationships were formed into a taxonomy of lexical relationship types. No claims to comprehensiveness were made; in fact, the opposite stance was emphasized, as specialty domains are bound to have unique relationships not appearing elsewhere.

Green argues quite succinctly that thesaural relationships are just as—if not more—important than conceptual entities, and that no current indexing language harnesses the theoretical power available in expressing relationships (1995a). In reviewing these so-called exhaustive typologies, Green maintains that there is no limit to the number and variety of relationships that might exist (1995b). Therefore, attempts to codify an exhaustive theoretical list are less valuable than analyzing which relationships actually do exist. Green and Bean (1995) subsequently attempted to determine which relationship types actually account for topical relevance. However, the only characteristic they found useful for retrieving relevant documents was that of contextual function, rather than any inherent properties of the relationships. While relevant for information retrieval, this work does not address potential value in relationships for browsing, navigation, and understanding a domain.

Attempting to improve subject access via controlled vocabularies, the Association for Library Collections & Technical Services' Subcommittee on Subject Relationships/Reference Structures spent nearly ten years exploring subject access structures, with a focus on related term references in LCSH (Miller, Olson & Layne, 2005). The Subcommittee identified 122 unique associative relationship types, although 41 of those were debated as hierarchical (Association for Library Collections & Technical Services, 1999). As with other previous authors, the Subcommittee deemed it unnecessary and unhelpful to determine, encode, and present such detailed relationships to users. However, the committee also admitted uncertainty regarding such a decision, especially in the context of users with varying needs and skills. Despite this potential, the focus of the committee shifted to discussion and display of pre-existing subject relationships, and has yet to return to the investigation of more specific semantic associative relationships.

# 3 Method

## 3.1 Study Design

While relationship terms and designators exist in a variety of knowledge organization systems, including classification schemes, indexing languages, and ontologies, one ubiquitous and accessible case was selected for this preliminary study. Library of Congress Subject Headings (LCSH), first created in 1898, were designed to describe a variety of library materials and cover the broad scope necessary to serve the Library of Congress as well as American academic and public libraries (Stone, 2000). In addition to American institutions, LCSH is used world-wide, on every continent except Antarctica, with 24 national libraries using LCSH in their national bibliographies (Stone, 2000). This scope and coverage, along with

the Subcommittee's previous lexical relationship work with LCSH, makes this thesaurus an ideal target for investigation.

To determine if potential library users can identify narrower distinctions of established associative relationship types, ground truth of relationship types first needed to be established. With such an extensive list from one of the most influential bodies in librarianship, the following six extant associative relationships recommended by the Subcommittee for systematic inclusion were chosen:

1. field of study/object of study
2. field of study/practitioner
3. agent/process
4. causal relationships
5. position in time and space
6. frequently interchangeable/near synonyms

To find example term pairs, the LCSH RDF data set from the Library of Congress (2012) was downloaded. Using Google Refine, the data set was distilled to display only terms with associative relationships (called "related terms" or RTs in LCSH). A cursory analysis showed a high number of languages and proper family names with RTs of the same type (e.g., "Khorezmi language RT Khorezmian Turkic language"; "Isaacson family RT Isacksson family"), which were eliminated from the sample since they would not be expressive of the relationship types chosen for review. From the remaining RTs, the authors sourced pairs that represented the six pre-identified relationships. The authors used their professional judgment and experience to select examples that were non-jargon and easily understandable to a layperson without further clarification or scope notes. Overly scientific and medical terms, such as genus-species or disease terminology (e.g. "Phytophthora infestans"), were eliminated. In addition to a need of specialized knowledge for comprehension, RTs including these types of terms often displayed specialized scientific relationships outside of the stipulated six. For each term pair, both authors had to 1) comprehend both terms in the pair without outside assistance (dictionaries, scope notes, etc.) and 2) agree on the relationship type being displayed. In this way, the authors strove to achieve "ground truth" answers. Five examples of each of the six relationship types were selected for inclusion in the survey (see Table 1).

| Relationship type | Term pairs |
|---|---|
| Field of study/object of study | Neurosciences AND Nervous system<br>Nuclear energy AND Nuclear engineering<br>Soil microbiology AND Soilborne plant diseases<br>Veterinary oncology AND Tumors in animals<br>Military psychiatry AND War neuroses |
| Field of study/practitioner | Taxonomists AND Biology—Classification<br>Criminal profilers AND Criminal behavior, Prediction of<br>Preventative medicine physicians AND Preventative medicine<br>Midwives AND Midwifery<br>Plastic surgeons AND Plastic surgery |
| Agent/process | Ear AND Hearing<br>Fermentation AND Leavening agents<br>Miracle workers AND Miracles<br>Eye AND Vision<br>Perspiration AND Sweat glands |
| Causal relationships | Plant diseases AND Crop losses<br>Distress in infants AND Crying in infants<br>Library overdues AND Library fines<br>Bacterial diseases AND Pathogenic bacteria<br>Inventory shortages AND Shoplifting |
| Position in time and space | American bison AND Buffalo meat<br>Grasses AND Hay<br>Spare parts AND Machine parts<br>Snow AND Meltwater<br>Lava AND Volcanic soils |
| Frequently interchangeable/ near synonyms | Kindness AND Benevolence<br>Window shades AND Blinds<br>Pillows AND Cushions<br>Engraving AND Etching<br>Pedicabs AND Rickshaws |

Table 1. Relationship types and example term pairs

After term pair selection, an online survey was deployed. The questionnaire included 30 multiple choice questions asking respondents to identify the relationship they perceived between the two concepts in the term pairs. The answers included the relationship types listed above, as well as a category for "other," in the event that they saw an alternative relationship beyond the suggestions provided. Data about level of library usage, level of familiarity with LCSH, and participants' opinions about whether they thought this level of granularity in relationships would be useful were also collected.

A total of 100 people were surveyed via Amazon Mechanical Turk (AMT), an established crowdsourcing platform frequently used to collect survey data (Figure 1). AMT has been successfully used for various tasks that require human intelligence, including a number of Natural Language Processing tasks (Snow et al., 2008), quality rating of Wikipedia articles (Kittur, Chi, & Suh, 2008), music mood and similarity judgment (Lee, 2010; Lee & Hu, 2012), and so on. On AMT, the task requester sets up a "HIT" (Human Intelligence Task) and human workers (called "Turkers") recruited by Amazon complete the HIT for a monetary reward. When the HIT is completed and submitted, the task requester reviews the HITs and approves or rejects them. Previous studies using AMT suggest that it is essential to have a filtering mechanism incorporated in the HIT to filter out bad responses. In addition to recruiting only participants with a HIT approval rate greater than 95%, we also included two randomly selected concept pairs in our survey to repeat in the HIT in order to check the consistency of user responses. Submissions where Turkers responded to the same question with different responses were rejected. As a result, of the 134 HITs submitted, 34 were rejected for this type of inconsistent answer. While we limited participants to those located in North America, we were unable to limit based on native language. The 100 responses required were collected in approximately 48 hours and the payment was $0.60 for completing each HIT. The average time Turkers spent on the HIT was 11 minutes and 28 seconds.

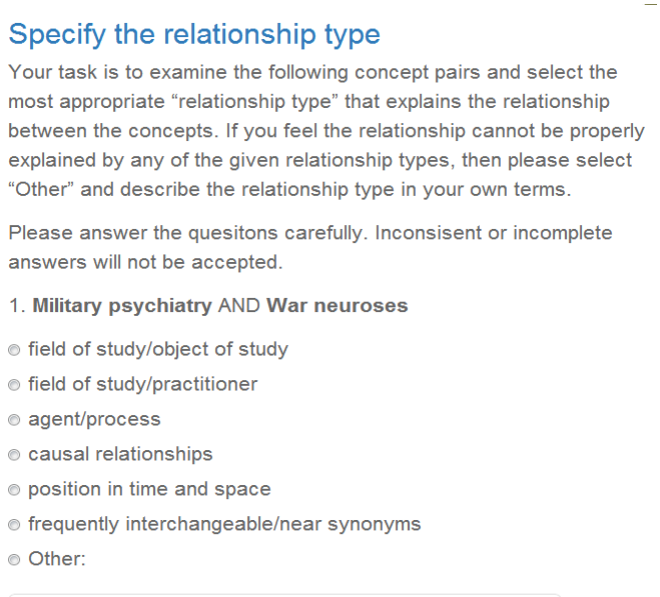


Figure 1. Screenshot of AMT HIT

## 3.2 Limitations

As a preliminary study, this research is but an introductory inquiry into conducting research about how users perceive associative thesaural relationships. While every attempt was made to use the most accurate methods possible, no study is without limitations. One major limitation occurred because of the transitive nature of associative relationships: if A is related to B, then B is related to A. Experienced vocabulary users and designers are aware of this transitive property. Therefore, when our experts selected term pairs to represent relationships, they looked for the existence of the relationship regardless of any directionality: for example, *Fermentation* is a process of *Leavening agents* was equated with the reverse relationship of leavening agents acting to ferment. However, when surveying participants, relationships were presented directionally, as there was no easy way to indicate this transitivity. Therefore, when presented with a term pair like "Fermentation AND Leavening agents" the choice of "agent/process" may not have been perceived as applicable due to the reverse direction of the relationship—"fermentation" is not an agent and "leavening agents" is not a process. Our investigation reveals that extreme care must be taken not to assume users' knowledge about thesaural relationships.

Future studies should be careful to take relationship directionality into account, perhaps by rotating the directionality of terms in a controlled fashion.

Another limitation of this study is the participant population. We acknowledge that workers on Amazon Mechanical Turk may or may not represent typical library patrons. Due to the preliminary nature of this study, ease of access to participant feedback was crucial both to refine the study design as well as offer a proof-of-concept with which to solicit future library settings. We attempted to control the sample somewhat by limiting participants to North America and only accepting participants with a successful track record of work. Additionally, we believe that all people—even those who have never used a library—may be considered potential library users. Library outreach, a major focus of many libraries, is purposefully designed to engage underserved as well as unserved users. Just because a participant has never used a library does not mean they do not have opinions about library-related concepts; library surveys that are limited to the current library population cannot represent these views. As libraries work to solicit library use to non-library users, it is possible that Turkers may be able to offer suggestions from that point of view.

# 4 Results

## 4.1 Perceptions of term relationships

For each RT pair, we examined how many respondents selected each of the seven answers (six relationship types plus "other"). The relationship type selected by the majority of respondents was identified, and is referred to as "agreement," meaning that the majority of respondents agreed on that answer. Agreement may be strong or weak: for a given RT pair, a vast majority of respondents may select one particular relationship type, while for another RT pair, respondents may differ widely in their selection of relationship type. For instance, the RT pair "Window shades AND Blinds" saw 93% of participants agreeing that *frequently interchangeable/near synonym* was the demonstrated relationship type. On the other hand, 62% of respondents said that "Plant diseases AND Crop losses" embodied a *causal* relationship, 20% said *agent/process* and 15% said *field of study/object of study*. For any given RT pair, an agreement of 80% and above was characterized as high agreement; agreement of 50%-80% as medium, and agreement of less than 50% as low. The distribution of high, medium, low agreement of each relationship type as represented by the term pairs was tallied (Table 2).

| | >80% agreement | 50-80% agreement | <50% agreement |
|---|---|---|---|
| Near synonym | 5 | 0 | 0 |
| Field of study/object of study | 4 | 1 | 0 |
| Causal | 0 | 3 | 2 |
| Field of study/practitioner | 0 | 3 | 2 |
| Agent/process | 1 | 3 | 1 |
| Position in time and space | 0 | 2 | 3 |

Table 2. Tally of term pairs at each level of agreement for the six relationship types

The highest overall level of agreement was for the *frequently interchangeable/near synonym* relationship type. For each of the five RT pair examples representing this relationship type, more than 80% of respondents identified the same relationship type as identified by the experts. The *field of study/object of study* relationship also showed high agreement overall, with only one term pair, "Nuclear energy and Nuclear engineering" seeing 68% agreement. *Causal* and *field of study/practitioner* relationships both showed medium to low agreement overall. The agreement about *agent/process* was varied: one RT pair, "Ear AND Hearing," showed high agreement (81%); three RT pairs ("Eye AND Vision", "Fermentation AND Leavening Agents", "Perspiration AND Sweat glands") showed medium agreement, and one RT pair ("Miracle workers AND Miracles") showed low agreement (48%). *Position in time and space* showed the lowest agreement overall.

While agreement represents respondents' perceptions, it does not necessarily align with the agreement from the ground truth. In the 10 total term pairs indicating high agreement (five *near synonym*, four *field of study/object of study*, and one *agent/process*), the agreed-upon response was indeed correctly identified. Of the remaining 20 pairs demonstrating medium or low agreement, 13 pairs were correct and seven pairs were not correct. Four of these incorrect agreements occurred when the correct relationship was *position in time and space* (two medium, two low) and three when the correct relationship was *causal* (one medium, two low). A contingency table (Table 3) illustrates the conflation. For the correct ground truth of *position in time and space* (y-axis), it is shown that participants most

commonly marked the incorrect answers of *near-synonymous/frequently interchangeable* and also *causal* relationships (x-axis). Likewise, when the correct answer was *causal*, participants commonly answered *near-synonymous* instead when answering incorrectly.

| | | Participants' answers | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | object of study/ field of study | object of study/ practitioner | agent/ process | causal | position in time and space | near syn./ freq. interchangeable | other |
| Correct answers | object of study/ field of study | 424 | 26 | 20 | 14 | 2 | 11 | 3 |
| | object of study/ practitioner | 48 | 289 | 95 | 16 | 5 | 22 | 25 |
| | agent/process | 30 | 43 | 316 | 66 | 4 | 25 | 16 |
| | causal | 72 | 2 | 50 | 236 | 8 | 128 | 4 |
| | position in time and space | 22 | 2 | 46 | 142 | 57 | 209 | 22 |
| | near syn./ freq. interchangeable | 3 | 1 | 9 | 38 | 4 | 438 | 7 |

Table 3. Contingency table comparing participants' answers against correct (ground truth) answers

## 4.2 Library experience

An overwhelming majority of respondents (88%) were library users who had never worked at a library. Three percent self-reported as library staff, but factoring in responses from former library staff raises this figure to 8%. One respondent (1%) self-identified as a library and/or information science researcher, and three (3%) offered other self-descriptions: "potential interest in library science studies," "computer geek" and "yogi." Approximately half of the participants (47%) had heard of Library of Congress Subject Headings but did not know much about them. Thirty-three percent had never heard of LCSH. Twenty percent reported some level of familiarity with LCSH, ranging from low to high.

The ratio of self-identified library users to librarians and other library employees (current and former) was 11:1 in the Amazon Mechanical Turk sample. Despite 88% of respondents self-identifying as library users, 29% of respondents said they had never used a library catalog. While library usage is certainly not limited to interaction with the catalog, the prevalence of digital materials and online access to library materials raises concern about accurate representation in the AMT sample. The relatively high percentage of non-catalog-users in this sample may reflect a larger target population than one focused on pre-existing patrons.

## 4.3 Opinions on the use of subject relationships

Participants were asked three open-ended questions: if they used the subject terms provided by the library catalog to find materials; if they thought it would be helpful if the library catalog displayed additional related term suggestions; and what types of term relationships or suggestions they thought would be most helpful. Space for additional comments was offered. Replies were coded for basic responses (yes/no/sometimes) and themes. To determine themes, each response was summarized and distilled into its main idea(s). These ideas were listed and similar ideas were consolidated and reworded. These reworded codes were then reapplied to the responses in order to see which themes emerged.

When asked about use of subject terms in library catalogs, a majority of the respondents (64%) said they did not use them. Nineteen percent replied that they did use subject terms to find library

materials. Seven percent of respondents said they used them sometimes, and 5% used to in the past but no longer do. Five percent did not respond.

Respondents who answered "yes" offered a variety of reasons for using subject terms in library catalogs. Some claimed to use them for quickness in finding desired materials. Some used them to find additional materials on the same subject and/or closely related subjects. Some used subject terms to narrow the scope of their query. Several respondents mentioned using subject terms when searching in a specific domain, including maps, English literature, art, gardening, and materials in a series. Three mentioned that while they had indeed used the subject terms, the results were unsatisfying: the terms were not specific enough, the experience was frustrating, and they could not find their desired item(s).

Most participants who did not use subject terms in library catalogs did not offer reasons or opinions. Of those who did, two major themes emerged. The first was a preference for other access methods. Many users preferred their own self-selected keywords over the controlled terms stipulated by the library.

> *When searching a library catalog, I usually pick my own search terms and ninety percent of the time come up with the book I'm looking for.* (P50)

Some preferred to search by title or author rather than subject, and one participant mentioned preferring interpersonal interaction over subject catalogs. The second major explanation for not using subject terms in library catalogs was that respondents found them unhelpful. Many noted that subject terms were too specific, and one person noted that s/he mostly searched for fiction, rendering the subject terms moot.

Despite almost two-thirds of respondents claiming that they did not use subject terms in library catalogs, 65% of respondents thought subject term relationships would be useful for library catalog users. Twelve percent felt unsure, several of whom indicated their ignorance about subject terms. Only 10% of respondents felt that subject term relationships would not be helpful in library catalogs. Thirteen percent did not respond.

While many participants indicated that additional subject term relationships would be "helpful," "save time" or increase "ease of use," they did not indicate how they thought these broad goals of efficiency and ease would come about. Others cited benefits such as exposure to relevant terms, materials, or areas of interest; narrowing or expanding search scope; suggestions for correct/authorized vocabulary terms; and confirmation of correct search path. Four participants simply responded that "more information is always better." Other responses included benefits for information literacy education and increased semantic web integration.

The minority of users who indicated that additional term relationships would not be useful offered more concrete reasons. Several participants indicated that since they had no problems with the current system, changes or additions were unnecessary. Some referred back to their answers that they did not use library catalogs or library subject terms. A few respondents mentioned a fear of a reduction in precision leading to distractions and "noise."

When asked specifically what kinds of subject relationships were likely to be helpful, participants offered a variety of suggestions. In attempt to determine which suggestions were most popular, the authors identified every relationship type mentioned in the qualitative responses. These types were organized and consolidated in a similar manner to the previous comments, and participant responses were coded according to these types (Table 4).

| Coded relationship type | Number of responses |
|---|---|
| Synonyms | 17 |
| Near-synonyms | 16 |
| Don't know | 15 |
| Causal | 10 |
| Field of study/object of study | 10 |
| Agent/process | 8 |
| None[1] | 8 |
| Closely related topics | 6 |
| Co-occurring search | 5 |
| Broader/narrower terms | 4 |
| Field of study/practitioner | 4 |
| Formal/vernacular | 4 |
| Related subjects | 3 |
| Related words | 3 |
| Opposites | 2 |
| Position in time and space | 2 |
| Time-specific | 2 |
| Channeler/spirit | 1 |
| Commonly associated terms | 1 |
| Culture-specific | 1 |
| Definitions | 1 |
| Item/function | 1 |
| Job/employment | 1 |
| Location-specific | 1 |
| Near-antonyms | 1 |
| Object/use | 1 |
| Predecessors | 1 |
| Related fields | 1 |
| Related titles | 1 |
| Same author | 1 |
| Similar subjects | 1 |
| Similar wording | 1 |
| Slightly related topics | 1 |

Table 4. Potentially helpful relationships suggested by participants

As with the question about helpfulness of additional subject relationships, some respondents offered vague suggestions, such as "Something that will make the search easier and less time consuming" (P25) or "…it would be nice to add some addition[sic] terms to make the search more specific and easier" (P70). Additionally, a handful of irrelevant responses, such as "that book[sic] should be classified in better categorical order" (P33) did not offer a germane answer to the question. Six vague and irrelevant responses could not be coded and were excluded from the tally.

# 5 Discussion

## 5.1 Identification and definition of relationships

Preliminary results demonstrate patterns of agreement—some stronger than others—indicating that potential library users can identify certain types of associative relationships. The high agreement and correct identification of the *near synonym* term pairs indicates ease of identification of this relationship type. The high agreement and correct interpretation of four out of five term pairs representing the *field of study/object of study* relationship also shows promise. The most common conflation for the fifth term pair (the one with only medium agreement), "Nuclear energy AND Nuclear engineering," was the *object of study/practitioner* relationship. Perhaps the speed of survey completion attributed to misreading "engineer" for "engineering," or perhaps respondents were simply confused about the true definition of

[1] Cases where participants specifically stated that no relationships would be useful (as opposed to a lack of response)

"practitioner," since the agreement level overall for *object of study/practitioner* was much lower than the *field of study/object of study* relationship which contained the same antecedent.

The *agent/process* relationship was unique in range of agreement. Eighty-one percent of respondents agreed that "Ear AND Hearing" represented the *agent/process* relationship, yet only 76% for the analogous "Eye AND Vision"; 60% for "Perspiration AND Sweat glands"; 51% for "Fermentation AND Leavening agents" (despite the word "agent" in the term pair); and 48% for "Miracle workers AND Miracles" (confused with *field of study/practitioner* by 40% of respondents).

*Position in time and space* appears to be a highly difficult relationship to identify. Agreement rates were low even when they indicated the intended relationship. In fact, *position in time and space* was only selected as a response (correct or not) 80 times in the entire survey. This may indicate a lack of ability to identify such a relationship or a lack of understanding about what the relationship type was intended to mean. This was also the relationship type that the researchers struggled to identify in LCSH and come to agreement with the examples.

It should be noted that the survey materials offered no definitions or explanations of either the terms within the pairs or the descriptions of the relationship types. This was an intentional design of the study, as no definitions were provided by the Subcommittee in the source material. Some respondents noted the lack of definitions and wanted further explanation and clarification. In the source material, the Subcommittee only provided a sample illustrative term pair for each relationship type. Even the authors found it difficult to articulate a definitive explanation for each relationship without biasing or otherwise influencing survey responses. Additionally, current displays of subject terms in library catalogs offer no context or explanation of terms, so excluding any specific relationship definition in the survey resembled a more realistic encounter. However, respondents' interest in and desire for such information may warrant further investigation into the integration of scope notes, definitions, and other explanatory context in user displays of library subject terms and relationships.

## 5.2 Synonymous and near-synonymous relationships

Synonymy, near-synonymy, and closely affiliated conceptions, such as affiliating a formal or technical term with its vernacular counterpart, was by far the common type of relationship perceived by users as potentially useful. The *near-synonymous/ frequently interchangeable* relationship was included as a possible survey answer, and thus it may be memorable to participants. However, other relationship types explicitly mentioned in the survey, such as *agent/process* and *field of study/practitioner*, were less often mentioned. *Near-synonymous* relationships were also the most often correctly identified. It seems clear that synonymous and near-synonymous relationships are generally easy to recognize and identify, and therefore be easy to cite as useful.

While the *near synonym/frequently interchangeable* relationship was considered by the Subcommittee to be an associative relationship, synonymous terms are generally considered equivalence relationships in thesaural construction (Aitchison, Gilchrist & Bawden, 2004), possibly due to ease of identification. However, users do not perceive differences in relationship types according to traditional thesaural definitions and categories. This leads to a question of where the line between synonyms and near-synonyms should be drawn. Respondents also mentioned usefulness of subject terms and relationships in helping to find "correct" search terms, i.e. authorized index terms:

> *It is sometimes difficult to arrive at the exact search term the catalog uses. Were it to display terms close to what I used, I might be able to find what I'm looking for more quickly.* (P88)

> *Definitely [useful], since a person may be thinking of a synonym or even a word closely related but not quite correct, and then have problems being led to the right books.* (P29)

Users are clearly aware of the existence of library thesauri and vocabularies, even if they do not understand how they work. They know they need to use a search term that matches the library index term to retrieve materials. Synonymous and near-synonymous relationships are useful to patrons because they help them navigate and access the authorized library vocabulary. Formal/vernacular relationships also play this role when leading a layperson to a technical term and vice versa.

> *I'd be most interested in figuring out how whatever I'm searching for is most commonly termed. For example, if I'm searching medical texts, I might have more success searching for "myocardial infarction" than "heart attack." I'd want a recommendation to suggest the former if I search for the latter.* (P91)

During thesaurus creation, designers decide to include technical or vernacular terms based on the domain of the thesaurus and its potential users. LCSH is designed to cover a broad scope, including

both academic and public library communities, where user needs vary widely. If one vocabulary is to serve them all, it may need to include more synonymous and near-synonymous relationships in order to assist users in retrieval. Near-synonymous terms can offer context, letting the user know he or she is on the right track:

> *[R]elated terms will confirm whether I'm searching for the right thing in the first place. It will also give me more ideas about what a certain topic is related to and what other people are looking for when they search for that term.* (P4)

An overview of a thesaurus also offers context and knowledge about a domain, but rarely do users have access to the entire thesaurus. Thesauri also use design considerations for contextual relevance, such as qualifiers, to distinguish "Mercury (planet)" from "Mercury (element)." Suggestions of near-synonymous or other related terms alongside the original subject term could offer additional useful context, helping to clarify vague, confusing, and imprecise terminology.

## 5.3 Defining relationships

A number of respondents specified "related" subjects, topics, or term as being useful relationships, but did not specify what kinds of relationships or how the terms or concepts should be related. The code "closely related topics" was only applied six times, but other codes that may overlap with this idea include related words (3); related subjects (3); commonly associated terms (1); related fields (1); related titles (1); similar subjects (1); similar titles (1); similar wording (1); slightly related topics (1). In terms of understanding relationship types, these codes, taken from participant language, are essentially meaningless. However, they reiterate the difficulty of articulating a definition for "related" terms and taxonomizing associative relationship types. In addition to "related" subjects, respondents also mentioned "similar" subjects. Our coding distinguished these as separate ideas, interpreting "similar" to mean a relationship containing some element of sameness between the two terms while "related" could also include relationships that did not include sameness, such as opposites or other non-similar relationships. This is, however, an interpretation by the researchers because it is difficult to know respondents' true intentions. It is also possible that participants themselves were not clear or able to articulate the concepts they had in mind. While the authors coded six responses as "closely related terms," one respondent mentioned both "closely related topics" and "slightly related topics":

> *Anything that can be closely or slightly related to the subject that I am looking for. It is up for me to decide what to look at.* (P95)

This reveals an interesting conceptualization of scale or measure. Are some topics more closely related than others? Are there ways of measuring these relationships? Perhaps such a measure could help untangle the confusion surrounding associative relationships by asking 'how much' are these two terms related rather than 'how' they are related.

Participants offered no definitions for their suggestions of useful relationship types. Suggestions such as "time-specific" and "culture-specific" relationships may lead to multiple interpretations. Do they mean a term in one culture that is synonymous for a term in another? Terms that change over time, such as the change from "Blacks" to "African Americans"? Without further clarification, it is impossible to clearly ascertain meaning. Future studies might benefit from asking for example term pairs along with relationship suggestions.

## 5.4 Co-occurrence

Several participants mentioned co-occurring search or subject terms as useful relationships:

> *It would great if it was similar to Amazon for shopping - people who searched for this te[r]m frequently searched for this term as well.* (P64)

Such relationships may be dynamic (such as co-occurring search terms that change over time as people's search queries change) or static (such as two subject terms that appear in the same bibliographic record). The reference specifically to Amazon may be because Mechanical Turk is an Amazon product and therefore at the forefront of respondents' minds or even visible on their screens while working. However, the prevalence of co-occurring relationships across the web demonstrates the influence of internet use on how people view library catalogs. The web offers many features that users find useful but have yet to be fully incorporated into library catalogs. Some catalogs, such as OCLC's WorldCat, do offer suggested subject terms based on concurrence in bibliographic records, but no respondents mentioned experiencing this in a library catalog.

## 6 Conclusions and Future Work

This preliminary study sought to determine whether users could identify narrower nuances of associative relationships and if they perceived such relationships as useful. Previous work regarding associative relationships assumed that narrower nuances would be unintelligible to users. The findings have shown that this is not necessarily the case. Some associative relationship types, especially *near-synonymous/ frequently interchangeable* and *object of study/field of study,* are easily identifiable by potential library users. Others, such as *position in time and space* remain problematic. However, there have been over 120 associative relationship types identified by various scholars, and study participants suggested many additional potentially useful relationship types. Potential library users do claim to perceive associative relationships as useful, but current low levels of subject use in library catalogs brings this perceived usefulness into question. Further work is needed to discern whether low levels of use are inherent to subject search and browse, or because of other design and implementation issues present in library catalogs.

Synonymous and near-synonymous relationships were most easily identifiable by respondents as well as frequently suggested as useful inclusions in library catalogs. Synonymous and near-synonymous relationships offer context for a subject term, help users identify authorized index terms, and assist users in navigating a vocabulary with extensive scope that covers professional and amateur domains. Future inclusion of more relationships of this type as well as the ability to display these relationships to users may offer substantial benefit, or may create information overload and favor recall over precision. Further study in this area is warranted. Other relationship types were more difficult to identify and define. Types of associative relationships may be less important than the closeness or proximity of the relationship. Future studies should examine the possibility of measuring degree of relatedness rather than taxonomizing it.

The use of a sample population from Amazon Mechanical Turk raises questions about the potential differences between library users and non-library users. The authors hope to undertake a parallel survey of real library users in order to compare with the respondents from Mechanical Turk. Future studies will also include more samples of each relationship type as well as other relationship types represented in LCSH beyond the six included in this study. However, these investigations are only a small step toward understanding of users' perceptions of associative subject relationships. Whether users would actually find associative relationships useful ultimately should be tested (1) with real user questions, (2) bibliographic resources retrieved by a source's subject heading or subject headings associatively related to that source heading, and (3) users' relevance judgments that are blind to which resources were retrieved by which heading.

While this study centered on library catalog subject terms in order to cover the broad scope necessary for a general audience, the issue of relationship types is relevant beyond libraries. Research shows that associative relationships are the most common types of relationships found in web search results (Milonas, 2012). Relationship recognition also offers implications for the internet at large, especially with regards to the semantic web. The purpose of the semantic web is to bring structure to web content that enables machines to parse sophisticated information (Berners-Lee, Handler & Lassila, 2001). This is made possible through "linked data"—a model that relies on relating entities through specifically defined relationships. However, machines are only as powerful as the human programming underlying them. While it may be easy to assert that Concept A is related to Concept B using linked data, it still does not shed light on the type of relationship that exists between those two concepts. On the other hand, detailed subject relationships such as those discussed in this paper may have powerful implications for machine understanding, if those relationships are clearly understood and defined. Imagine the inferences possible if we could stipulate cause and effect or processing agents. A more in-depth understanding and specific expressions of subject relationships may offer advances beyond library catalogs to the world at large.

## References

Aitchison, J., Gilchrist, A., and Bawden, D. (2004). *Thesaurus construction and use: a practical manual*. New York; London: Europa Publications.

Association for Library Collections & Technical Services, Subcommittee on Subject Relationships/Reference Structures. (1999). Final Report to the ALCTS/CCS Subject Analysis Committee. Retrieved August 26, 2014, from http://www.ala.org/alcts/mgrps/camms/cmtes/sac/inact/subjrel

Berners-Lee, T., Hendler, J. and Lassila, O. (2001). "The Semantic Web." Scientific American, 284(5): 29-37.

Green, R. (1995a). Syntagmatic relationships in index languages: a reassessment. *Library Quarterly, 65*, 365-385.

Green, R. (1995b). Topical Relevance Relationships. I. Why Topic Matching Fails. *Journal of the American Society for Information Science, 46*(9), 646-653.

Green, R., and Bean, C. A. (1995). Topical Relevance Relationships. II. An Exploratory Study and Preliminary Typology. *Journal of the American Society for Information Science, 46*(9), 654-662.

Kittur, A., Chi, E. H., Suh, B. (2008). Crowdsourcing User Studies with Mechanical Turk. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 453-456.

Lee, J. H. (2010). Crowdsourcing Music Similarity Judgments using Mechanical Turk. Proceedings of the 11th International Society for Music Information Retrieval Conference, 183-188.

Lee, J. H. and Hu, X. (2012). Generating Ground Truth for Music Mood Classification using Mechanical Turk. Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries.

Library of Congress. (2012). Authorities & Vocabularies. Retrieved March 6, 2012 from http://id.loc.gov/

Miller, D., Olson, T., and Layne, S. S. (2005). Promoting Research and Best Practices in Subject Reference Structures: A Decade of Work by the Subject Analysis Committee. *Library Resources & Technical Services, 49*(3), 154-166.

Milonas, E. (2012). Classifying Web term relationships: an examination of the search result pages of two major search engines. In Neelameghan, A. and Raghavan, K.S. (eds.), Categories, Contexts and Relations in Knowledge Organization: Proceedings of the Twelfth international ISKO Conference, 6-9 August 2012, Mysore, India. Würzburg, Germany, Ergon-Verlag: 201-205.

Neelameghan, A., Maitra, R., and International Federation for Documentation. (1978). *Non-hierarchical associative relationships among concepts: identification and typology*. Bangalore: Documentation Research and Training Centre, Indian Statistical Institute.

Nutter, J. T. (1989). *A lexical relation hierarchy*. [Blacksburg]: Dept. of Computer Science, Virginia Polytechnic Institute and State University.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A.Y. (2008). Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 254-263.

Soergel, D. (1974). Indexing languages and thesauri: construction and maintenance. Los Angeles: Melville.

Stone, A. T. (2000). The LCSH century: a brief history of the Library of Congress subject headings, and introduction to the centennial essays. *Cataloging & Classification Quarterly, 29*(1/2), 1-15.

Willetts, M. (1975). An Investigation of the Nature of the Relation between Terms in Thesauri. *Journal of Documentation, 31*(3), 158-184.

## Table of Figures



## Table of Tables